Real-time Upper Body Detection and 3D Pose Estimation in Monoscopic Images

Antonio S. Micilotta, Eng-Jon Ong, and Richard Bowden

Centre for Vision, Speech and Signal Processing,
University of Surrey,
Guildford GU2 7XH, Surrey, United Kingdom.
{e.ong,r.bowden}@surrey.ac.uk

Abstract. This paper presents a novel solution to the difficult task of both detecting and estimating the 3D pose of humans in monoscopic images. The approach consists of two parts. Firstly the location of a human is identified by a probabalistic assembly of detected body parts. Detectors for the face, torso and hands are learnt using adaBoost. A pose likliehood is then obtained using an a priori mixture model on body configuration and possible configurations assembled from available evidence using RANSAC. Once a human has been detected, the location is used to initialise a matching algorithm which matches the silhouette and edge map of a subject with a 3D model. This is done efficiently using chamfer matching, integral images and pose estimation from the initial detection stage. We demonstrate the application of the approach to large, cluttered natural images and at near framerate operation (16fps) on lower resolution video streams.

1 Introduction

Our objective is to automatically locate the presence of human figures in natural images, and to estimate the 3D skeletal pose of that figure. Fitting a 3D model to a monocular image of a person requires a reliable estimate of the position of that person. Our first objective is therefore to robustly estimate the location and approximate 2D pose of a user in a real world cluttered scene. This is a challenging task as the shape and appearance of the human figure is highly variable. We have extended AdaBoost [15] to create body part detectors for the face, torso and hands. Detections are then assembled into an upper body pose via RANSAC [4] in real-time. Once an upper body 2D pose is selected, the second objective, is to reconstruct the 3D upper body pose making use of a prior dataset of human motion capture.

Human detection is often facilitated by detecting individual body parts, and assembling them into a human figure. Ioffe and Forsyth [6] make use of a parallel edge segment detector to locate body parts, and assemble them into a 'body plan' using a pre-defined top level classifier. Similarly, Felzenszwalb and Huttenlocher [3] use rectangular colour-based part detectors, and assemble detected parts into a body plan using pictorial structures. Ronfard et al.[10] use detectors trained by dedicated Support Vector Machines (SVM) where a feature set consists of a Gaussian filter image and 1st and 2nd derivatives. Haar wavelets are used by Mohan et al. [9] to represent candidate regions and SVMs to classify the patterns. Roberts et al. [11] have created probabilistic region templates for the head, torso and limbs where likelihood ratios for individual parts are learned from the dissimilarity of the foreground and adjacent background distributions. Mikolajczyk et al. [8] model humans as flexible combinations of boosted face, torso and leg detectors. Parts are represented by the co-occurrence of orientation features based on 1st and 2nd derivatives. The procedure is computationally expensive, and 'robust part detection is the key to the approach' [8].

Our approach is novel in that it uses RANSAC to combine appearance, colour and structural cues with a strong prior on pose configuration to detect human structures. 3D reconstruction from a single camera has also recieved considerable attention. Howe [5] et al. tracked 20 body points from a monocular sequence, and adopted a bayesian framework to compute prior probabilities of 3D motions with the aid of training data. An alternative is proposed by Sigal et al. [13] where the human body is represented as a graphical model where relationships between body parts are represented by conditional probability distributions. The pose estimation problem becomes one of probabilistic inference over a graphical model with random variables modelling individual limb parameters. Fitting a 3D model to a single image of an object is achieved by comparing shape and edge templates of an example database to the object of interest. This has been applied to hand pose estimation [14] where shape matching follows a cascaded approach to reduce the number of edge template comparisons. Most 3D reconstruction approaches rely upon tracking assuming an initial pose is already known. Here, we combine robust detection with 3D estimation allowing the visually accurate reconstruction of pose within a single image. We also extend this approach to tracking in a video stream.

This paper is set out as follows: A basic discussion of AdaBoost applied to object detection is presented in Section 2. Our first contribution offers a method of assembling body part detections using RANSAC, a heuristic, and an a priori mixture model of upper-body configurations (Section 3). The chosen assembly is then used to assist in reconstructing the corresponding upper body 3D pose (Section 5). Section 5.1 describes the acquisition of the database of 2D upper body frontal poses from the 3D animated avatar, which is then subdivided into subsidiary databases. Matching the silhouette and edge templates of the user to those of example databases is discussed in Section 5.4.Finally, results are shown, and conclusions drawn.

2 Boosted Body Parts Detectors

Boosting is a general method that can be used for improving the accuracy of a given learning algorithm.More specifically, it is based on the principle that a highly accurate or 'strong' classifier can be produced through the linear combination of many inaccurate or 'weak' classifiers. The efficiency of the final classifier

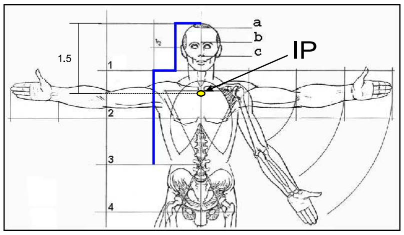


Fig. 1. Virtuvian Man

is increased further by organising the weak classifiers into a collection of cascaded layers. This design consists of a set of layers with an increasing number of weak classifiers, where each layer acts as a non-body-part rejector with increasing complexity. An input image is first passed to the simplest top layer for consideration, and is only moved to the next layer if it is classified as true by the current layer. The reader is directed to [15] for a detailed discussion of AdaBoost cascades.

Using AdaBoost, we separately trained four different body part detectors using their respective image databases. In order to detect a specific body part in a bounding box, we offset all the weak classifiers belonging to that detector to that location. A positive or negative detection is then computed by combining weak classifier outputs in strong-classifier layers. Each detector returns a score for part detection, which is then normalised to produce a likelihood, defined as L_F, L_T, and L_H respectively.

Since detections are performed in gray scale, it would be advantageous to exploit colour cues to contribute to a detection's legitimacy. Here, the face and hands benefit from this constraint. Initially, a weak skin colour model in the Hue-Saturation colour space built from a large selection of natural images containing skin regions. Using this generic skin model, we determine the median skin likelihood for the face (L_{FS}) and from this face detection we obtain a refined user specific skin model for use in hand detection (L_{HS}).

3 Human Body Assembly

The methods described in the previous sections provide the detected body parts needed to construct a human model. To ensure that most of the body parts are detected, fewer layers in the cascade are selected, resulting in a larger number of false detections. In order to determine liklely body configuration from the numerous detected body parts, a three step process is followed: 1) RANSAC is used to assemble random body configurations, each consisting of a head, a torso, and a pair of hands. A weak heuristic is then applied to each configuration to eliminate obvious outliers (3.1). 2) Each remaining configuration is compared to an a priori mixture model of upper-body configurations, yielding a likelihood for the upper body pose (3.2). 3) A resultant likelihood for each configuration is obtained by combining the likelihood determined by the prior model with those of the body part detectors and corresponding skin colour (if applicable). Configurations with a high likelihood are determined and the support assessed via RANSAC (3.3).

3.1 Building a coarse heuristic

An image with several human figures and dense background clutter can produce multiple part detections in addition to false detections. RANSAC selects subsets of detections that represent body configurations, however testing all these configurations would be computationally expensive; a coarse heuristic is therefore employed to discard unlikely configurations.

Rules of the heuristic are designed according to a generic human model, and include a reference length measurement. Referring to Da Vinci's Virtuvian Man (Figure 1) the human figure is subdivided into eight lengths, each equal to the "head length (the top of the skull to the chin). For the purpose of this paper, this length is referred to as a *skeletal unit length*. The head can be further subdivided into 3 lengths, a,b and c – a typical face detection occupies b and c, thereby allowing us to approximate the skeletal unit length.

Comparing the ROC curves of Figure 6a it is evident that the face detector is the most robust. For this reason, the face detector forms the base for every body configuration. The skeletal unit length and centre position of a selected face is determined, and form the parameters that assist in solving a body configuration.

The rules of the heuristic are set out in the following order, with x and y referring to horizontal and vertical directions: 1) A torso is added to the model only if: its centre x position lies within the face width; the torso scale is approximately $3 \times$ face scale ($\pm$ 0.5); the face centre lies within the detected torso region. 2) A pair of hands are added only if: both hands are less that $4 \times$ skeletal unit lengths from the face; the hand scale $\approx$ face scale ($\pm$ 0.2). False hand detections form the bottleneck in the system as a large number are accepted by the heuristic. The configurations that are passed by the heuristic are then compared to an a priori mixture model of upper-body configurations to obtain a likelihood for the upper body pose (see equation 1), which plays an important role in eliminating false hand detections as awkward hand poses yield a low likelihood.

3.2 Prior Data for Pose Likelihood

In this second step, we use an a priori mixture model of upper-body configurations to estimate the optimal upper body pose. Each body configuration obtained by the above-mentioned selection process provides the position of 8 points, namely the four corners of the torso detector, the chin and brow of the face detector, and the hands. These 8 x, y coordinates are concatenated to form a feature vector $\mathbf{Y} \in \Re^{16}$.

An a priori model ϕ of upper-body configurations was built from approximately 4500 hand labelled representative examples ($\in \Re^{16}$ as above) from image sequences of subjects performing various articulated motions. A Gaussian Mixture Model (GMM) is then used to represent this non-linear training set. The number of components k is chosen through analysis of the cost function, constructed from k-means. Here, $k = 100$. k 16x16 covariance matrices $Cov_{\phi,k}$ are formed from data set ϕ, where $Cov_{\phi,k} = \frac{1}{N_k - 1}(\phi_i - \mu_{\phi,k})(\phi_i - \mu_{\phi,k})^T$, and $\mu_{\phi,k}$ is the mean of each component of the GMM. A measure of how well each newly assembled body configuration fits the prior data set can now be determined.

The Mahalanobis distance between the configuration and the prior is determined and a final pose likelihood L_P is obtained from the weighted sum of the likelihoods for each component:

$$L_P = \sum_{i=1}^{k} \frac{N_i}{N} \left[\left(2\pi^{\frac{d}{2}} |Cov_{\phi,i}|^{\frac{1}{2}} \right)^{-1} exp(-\frac{1}{2} md^2_{\phi,i}) \right] \tag{1}$$

3.3 Final Configuration Selection

The eight determined likelihoods, namely the mixture model (L_P), face (L_F), face skin (L_{FS}), torso (L_T), left hand (L_{LH}), left hand skin (L_{LHS}), right hand (L_{RH}) and right hand skin (L_{RHS}) are combined to provide an overall body configuration likelihood, L_{BC}.

$$L_{BCi} = L_{Pi}.L_{Fi}.L_{FSi}.L_{Ti}.L_{LHi}.L_{LHSi}.L_{RHi}.L_{RHSi} \tag{2}$$

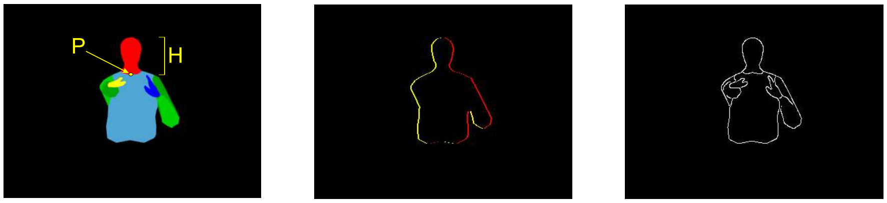
To determine the most likely pose consensus for a specific pose is accumulated by RANSAC. This is possible as objects tend to produce multiple overlapping detections.

4 Detection in sequences

Extending this work to video sequences allows us to take advantage of background segmentation and to apply the detectors in a tracking framework.

Our background removal algorithm was originally developed for exterior visual surveillance and relies upon modelling the colour distribution with a Gaussian mixture model on a per pixel basis [7]. This allows each pixel to be assigned a foreground likelihood which increases according to sudden intensity variation. We apply the detectors on the full natural frame, and include the mean foreground likelihood L_{FG} of a detection's bounding box. The body configuration likelihood of Equation 2 is therefore updated as follows:

$$\begin{aligned} L_{BC_i} &= (L_{Pi}) \times (L_{Fi}.L_{FSi}.L_{FG_{Fi}}) \times (L_{Ti}.L_{FG_{Ti}}) \\ &\times (L_{LHi}.L_{LHSi}.L_{FG_{LHi}}) \times (L_{RHi}.L_{RHSi}.L_{FG_{RHi}}) \end{aligned} \tag{3}$$



(a) (b) (c)

Fig. 2. (a) Frontal 2D representation of 3D model (b) Boundary image (c) Edge map

The chief advantage of detection in a video sequence lies in the tracking framework where the search space is localised in subsequent frames, thereby reducing the number of false detections, the number of hypotheses assessed by RANSAC, and therefore improving speed performance. An initial face detection is conducted as before, with consequent body part detections limited by the heuristic proximity rules as defined in section 3.1. Subsequent position and scale variations of each detector are governed by prior detections. Should a body part fail to be detected, the search region for the corresponding detector is increased linearly and the scale is adjusted by a Gaussian drift term until the detector recovers.

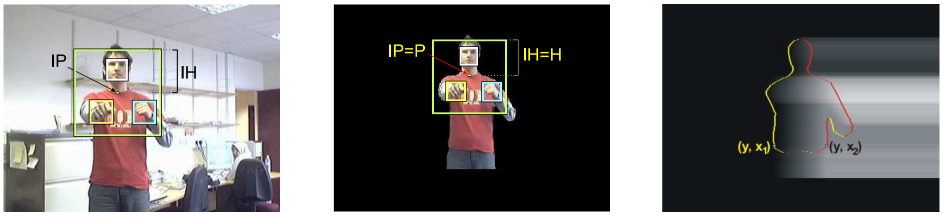
5 Estimating the 3D pose

Once an upper body assembly is selected, we estimate the corresponding 3D pose by matching the silhouette and edge map of the user to those of the animated 3D avatar.

5.1 Data acquisition

Using a 3D graphics package, a skeleton is skinned with a generic human mesh to resemble a person wearing loose fitting clothing and rendered using cell shading. A rendered model with one colour level resembles a simple silhouette. We therefore colour the respective body parts independently to preserve edges between different limbs and the body. The left and right hands are coloured blue and yellow respectively to provide independent labelling. Only the upper body is rendered by assigning the lower body a transparent material.

A single target camera (a camera whereby the camera-to-target distance remains fixed) is then attached to the chest bone of the skeleton, and is allowed to roll in accordance with it. The skeleton is then animated and rendered with a variety of movements using motion capture data (5000 frames), yielding a database of 2D frontal view images (Frontal View Database) of an upright upper body that has a fixed scale, and is centred at position P (Figure 2 (a)).



(a) (b) (c)

Fig. 3. Input Image: (a) Original (b) Adjusted (c) Integral image / boundary overlap

Subsidiary datasets The images of the Frontal View Database are then used to produce a hierarchy of three subsidiary databases. These are computed offline, and are loaded in when the application is executed. All examples in these databases are indexed according to the original frontal image database and the corresponding pose configuration data that generated it. From parent down: **1) Hand Position Database.** This consists of the 2D positions of the left and right hands that are obtained by determining the centroid of the blue and yellow (hand) regions of each frame. **2) Silhouette Database.** This is easy to create as the background of each example is black. The boundary of silhouette images are efficiently stored as entry and exit pairs for each row of the silhouette. This representation also offers a fast and efficient method of comparison to the input silhouette, which is represented as an integral image (see Section 5.4). **3) Edge Map Database**. Conducting an edge detection on the cell shaded and multi-coloured model provides clean edge images (Figure 2 (c)). Again, to conserve memory, only the edge locations are stored.

5.2 Input image adjustment

The sections below discuss the processes that occur at run-time, after the subsidiary databases have been loaded. Referring to an example of the Frontal View Database (Figure 2 (a)), the length from the top of the head to the neckline H, is constant across all examples, and is used as the reference point with which to scale the input image. Position P and length H are pre-computed.

Comparing the Frontal View Database and its subsidiaries to the input image requires that the input image foreground exists in same spatial domain (see Figure 3 (b)). To do this, the input image neck centre IP and head length IH must be determined. The assembled body determined in Section 4 provides the dimensions of the face, from which the skeletal unit length is approximated (Section 3.1).

The scale factor is determined by $S = IH/H$, and the offset from P to IP is determined by $offset = P - IP/S$. The input image is scaled and translated in a single pass, creating the *adjusted input image* (*AdjIm*) of Figure 3 (b). We then extract an input silhouette IS and edge map from this adjusted input image.

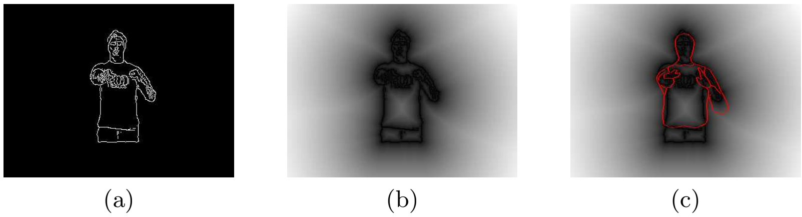


Fig. 4. (a) Edge image (b) Distance image (c) Chamfer match

5.3 Extracting subsidiary database examples

Before conducting silhouette matching, we initially extract a subset of the Silhouette Database by considering the user's hand positions. Using the left and right hand bounding boxes provided by the tracking algorithm as reference, we search through the Hand Position Database for hand positions that are simultaneously contained by these bounding boxes, and extract the corresponding examples from the Silhouette Database. This too can be precomputed by indexing examples in the database to the gaussian components of the GMM used in the pose likliehood. From the possible examples identified; a matching score is therefore calculated for each example as per Section 5.4.

5.4 Silhoutte Matching using Integral Images and Chamfer Matching

We determine a set of matching scores for the Slihouette Database subset by computing the percentage pixel overlap between the input silhouette and each example. The matching procedure is made more efficient by using an integral image II as an intermediate representation of the input silhouette IS.

The II encodes the shape of the object by computing the summation of pixels on a row by row basis. The value of the $II(x, y)$ equals the sum of all the non-zero pixels to the left of, and including $IS(x, y)$:$II(x, y) = \int\limits_{i=0}^{x} IS(i, y)\mathrm{d}i$

The entire II can be computed in this manner for all (x, y), however for efficiency we compute this incrementally: $\forall x, y\ II(x, y) = IS(x, y) + II(x-1, y)$ Figure 3 (c) offers a visualisation of the integral image of the input silhouette (extracted from Figure 3 (b)), with a silhouette boundary example of the Silhouette Database superimposed. Referring to Figure 3 (c), the number of pixels between boundary pair (y, x_1) to (y, x_2) is computed as $N_B(y) = x_2 - x_1 + 1$. The number of pixels of the input silhouette for the corresponding range is therefore computed as $N_{IS}(y) = II(y, x_2) - II(y, x_1) + 1$,where $\sum N_B$ and $\sum N_{IS}$ are computed for all boundary pairs, and the matching score is therefore computed as $S = \sum N_{IS} / \sum N_B$. This score is computed in a few hundred operations; considerably less than tens of thousands of pixel-pixel comparisons.

A matching score is computed for each example of the Silhouette Database subset, the top 10% of which are compared to the corresponding edge maps from the Edge Map Database using Chamfer Matching [1]. To achieve this, the distance transform [2] of the input edge image (Figure 4 (a)) is obtained to 'blur' the edges (Figure 4 (b)), where the intensity of a distance transform pixel is proportional to its distance to an edge. We then superimpose the example edge map on the distance image, and determine the *edge distance* – the mean of the distance image pixel values that co-occur with example edge maps. The example that yields the shortest distance represents the best match, and is used to access the 3D data from the original database.

6 Results

Comparison of the different part detectors is a difficult task. The most obvious problem is that each part is of different scale, and we would therefore expect a larger number of false hand detections than false torso detections for example. Our in-house face database consists of colour images containing 500 faces, and is similar in size to the MIT-CMU face database (507 faces). The torso were tested on 460 (of 900) images of the MIT pedestrian database, while the hand detector was tested on a colour image database containing 400 hands. Figure 6a shows the detection performance of the detectors applied to their respective test datasets, where layers from the classifier are removed to increase the detection rate. In this research, detection is considered true if at least 75% of its bounding box encloses the groundtruthed body part. In addition, we do not merge overlapping false detections as in [12]. We have plotted two curves for the face detector to show the advantage of including colour. The face detector proves to be the most robust of the detectors, since the face is a self contained region. Other body parts are affected by background clutter and have a greater variability in appearance. Due to the high variability of hand shape, we expect the hand detector to offer the poorest performance.

Making use of the ROC curves plotted for each detector, the desired number of layers was chosen such that the probability of detecting all objects was no less than 80%, with the trade-off of an increased number of false detections. The initial detections from the body part detectors are rapidly eliminated using RANSAC and the heuristic, before being narrowed down to the body configuration with the largest likelihood as determined by the joint-likelihood model as shown in Figure 5 (top row). The entire process from detection to assembly takes approximately 5 seconds on a P4, an improvement over [8], which takes 10 seconds and does not include hand detection.

The middle row of Figure 5 illustrates the body part assembly of a subject walking into an office and performing hand gestures using background segmentation as described in Section 4. The scene is particularly complex with wooden furniture and cream walls, thereby yielding poor background segmentation. Our assembly system overcomes these difficulties and operates at 8 frames/sec (frames sized at 640x480), a considerable improvement from the static image case. For

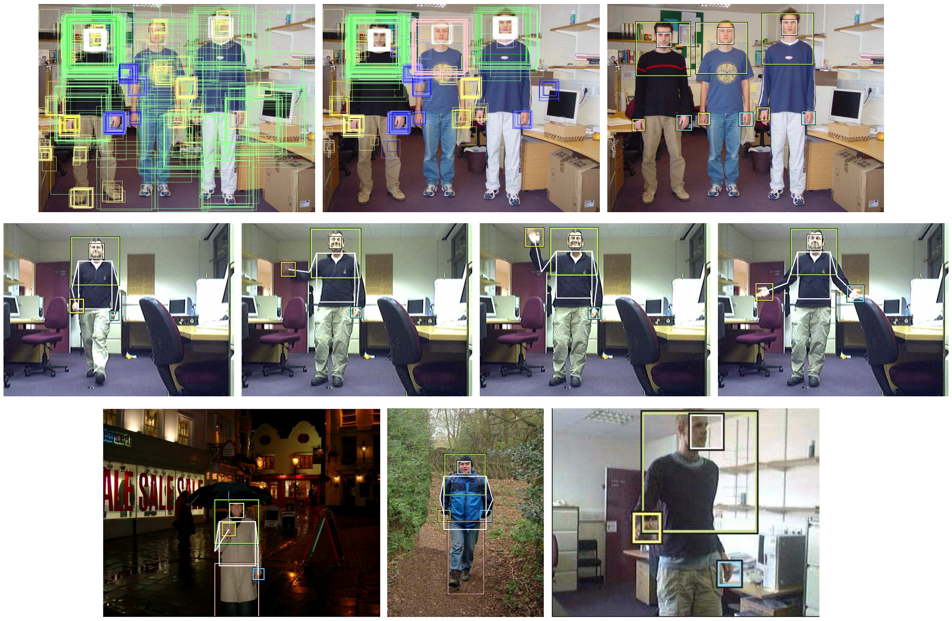
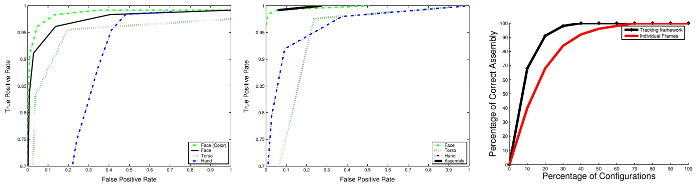


Fig. 5. Top row (from left): All detections, Reduced detections and Final Assembly. Middle Row: Body part assembly from a video sequence. Bottom row (from left): synthesised leg positions due to leg detection failure, synthesised hand positions due to hand occlusions, detections for non-frontal body and face poses.

completeness, elbow positions that have been determined by statistical inference [7] are given. A corresponding performance curve for this sequence is given (Figure 6b). To maintain consistency with the performance curves of Figure 6a, each frame of this sequence was treated as a discrete image, with the search space encompassing the entire image. However, to illustrate the benefit of background suppression, the hand detector includes the foreground fitness, and offers similar performance to the torso detector. In using a sequence the performance of the assembly method on a full subject could be evaluated. As expected, the assembly curve supersedes the others, illustrating the robust false part elimination of the assembly methodology.

To test our method for tolerance to occlusions, an increasing number of body parts detections were deliberately removed randomly at each frame. The number of correct assembly body configurations found across the entire video sequence was calculated, repeated 5 times and the mean result of correct assembly vs percentage of removed body parts obtained (Figure 6c). The black plot is the output from using a tracking framework where the detection window for each part is limited. The red plot treats each frame independently and has lower performance due to increased ambiguities. Also illustrated is how other cases of occlusions and non-frontal body poses through synthesis of missing parts are handled (Fig-



(a) (b) (c)

Fig. 6. Detector performance on (a) test databases (b) video sequence (c) simulated occlusions of body parts.

ure 5(bottom row)). Figure 7 shows the selected body assembly of subjects from various sequences and its representative CG model. The frames are captured at 320x240, and runs at 16 frames/sec. Comparison of the various scenes shows the matching method to be invariant to the user's scale and position.

7 Conclusions

We have extended an existing boosting technique for face detection to build two additional body part detectors. Due to the variability of these body parts, their detection performance is lower, and a technique was developed to eliminate false detections. By combining a coarse body configuration heuristic with RANSAC and an a priori mixture model of upper-body configurations, we are able to assemble detections into accurate configurations to estimate the upper body pose. When this approach is applied to a video sequence, exploitation of temporal data reduces the false detection rate of all the detectors, and improves speed performance dramatically. We have also been successful in matching a corresponding 3D model to the selected body part assembly. Matching by example does however require a large example dataset, and we have therefore stored our datasets in their simplest forms. These simple representations Examples from the large example dataset, were stored in their simplest forms, for fast access, contributing efficiency to the fast matching methods employed. Furthermore, the hierarchical structure restricts analysis to subsets of the subsidiary databases, thereby contributing to the real-time aspect of the approach.

Acknowledgements

This work is partly funded under the EU FP6 Project "COSPAL", IST-2003-2.3.2.4.

References

1. H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proc. of Joint Conf. Artificial Intelligence*, pages 659–663, 1977.

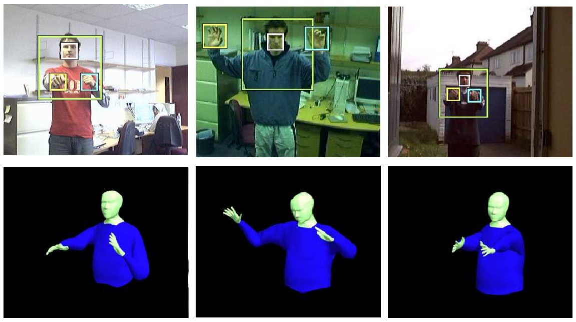


Fig. 7. Frontal pose with corresponding 3D model

2. P. Felzenszwalb and D. Hurrenlocher. Distance transforms of sampled functions. Technical Report TR2004-1963, Cornell Computing and Information Science, 2004.
3. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient matching of pictorial structures. In *Proc. of CVPR*, volume 2, pages 66 – 73, 2000.
4. M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In *Comm. of the ACM*, volume 24, pages 381–395, 1981.
5. N. Howe, M. Leventon, and W. Freeman. Bayesian reconstruction of 3d human motion from single camera video. In *Advances in Neural Information Processing Systems*, volume 12, pages 820–826, 2000.
6. S. Ioffe and D. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001.
7. A.S. Micilotta and R. Bowden. View-based location and tracking of body parts for visual interaction. In *Proc. of British Machine Vision Conference*, volume 2, pages 849–858, September 2004.
8. K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust body part detectors. In *Proc. of ECCV*, volume 1, pages 69–82, 2004.
9. A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on PAMI*, 23(4):349–361, April 2001.
10. B. Triggs R. Ronfard, C. Schmid. Learning to parse pictures of people. In *Proc. of ECCV*, volume 4, pages 700–707, 2002.
11. T. Roberts, S. McKenna, and I. Ricketts. Human pose estimation using learnt probabilistic region similarities and partial configurations. In *Proc. of ECCV*, pages 291–303, 2004.
12. H.A. Rowley, S. Baluja, and T.Kanade. Neural network-based face detection. *IEEE Transactions on PAMI*, 20(1):23–38, January 1998.
13. L. Sigal, M. Isard, B. Sigelman, and M. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. In *Proc. of Advances in Neural Information Processing Systems*, volume 16, pages 1539–1546, 2003.
14. B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Hand pose estimation using hierarchical detection. In *Workshop on Human Computer Interaction*, pages 105–116, 2004.
15. P. Viola and M. Jones. Robust real-time object detection. In *Proc. of IEEE Workshop on Statistical and Computational Theories of Vision*, 2001.